Nathan Storey

A Review & Critique of “Evaluation Study of the ClassPace Program in Fenton, Michigan”

*Review*

Cook and his colleagues at the Center for Research and Reform in Education (CRRE) were contracted by Odysseyware to conduct an independent evaluation of the Odysseyware ClassPace program during the 2018-2019 academic school year. The ClassPace program is an “online learning environment…that provides supplemental digital curricula for grades 3-12 in math, language arts, science, and social studies…allow[ing] teachers to differentiate instruction in blended learning classrooms” Cook, et al., 2019). Their study sought to identify any relationship between participation in the program and students’ math achievement and attitudes towards math, as well as teachers’ perceptions of the program. To accomplish this, the researchers conducted their study in Fenton Area Public Schools in Michigan, a district of 3,400 students in three elementary schools, one middle, and one high school. Their analysis included 189 fifth-grade students (123 in their treatment group, 66 in their control group), with two teachers across two different elementary schools in the treatment group, as well as three teachers in the third district elementary school as their comparison group.

Data collection included two rounds of teacher interviews and student surveys (serving as pre- and post-test datapoints). Teachers were interviewed related to their opinions of using ClassPace, any changes they had made to their teaching practices, and any impacts on student academic achievement related to ClassPace. Meanwhile, the student surveys concerned student attitudes toward mathematics. Other data collected included ClassPace usage data, and NWEA Measures of Academic Progress (MAP) standardized mathematics exam results from winter 2018 and spring 2019.

The researchers failed to find any statistically significant connection to math scores based on enrollment in ClassPace, but did find a significant positive relationship between math achievement and increased usage of ClassPace. Roughly each additional lesson completed was associated with about 33% of a point on the spring 2019 MAP test. Cook, et al., also found small positive gains in student attitudes regarding math enjoyment and understanding. Finally, they noted that teachers responded positively to ClassPace in relation to their teaching as well as student learning.

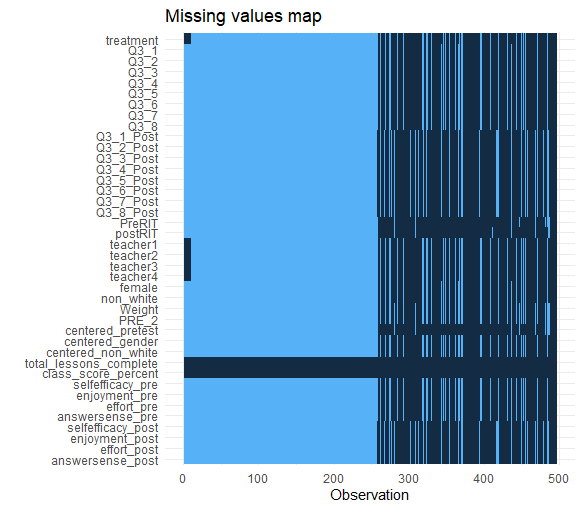
*Critique*

This is a relatively straightforward program evaluation. The authors do not delve into the theoretical background for the ClassPace program or the reason it is worthy of study. CRRE was contracted by Odysseyware, the company that administers ClassPace, which was all the rationale they needed, which is natural. ClassPace appears to be based on common theoretical pedagogical processes (“I do, we do, you do”), including steps of setting of clear learning objectives, clear instruction, student practice with guidance, student practice on their own, and final review of learning (LaMalfa, 2018). The blended approach is intended to be flexible in order to adapt to multiple learning styles (ibid). None of this is described by the study, which is focused instead on the evaluation approach of the study as a contracted program evaluation.

Peer-reviewed studies of ClassPace do not seem commonly available currently, but Odysseyware did publish several case studies on their website, all showing large increases. One in a middle school in Little Rock, Arkansas suggested students experienced a 75% improvement on assessments, as well as large gains on homework and classroom behavior (LaMalfa, 2018). Given the biased nature of the publications, and the lack of concrete information or research methodology description backing up the results shared in the case studies, it does not seem surprising that the results of this study did not match the previously shared results. However, the study has several limitations worth considering as part of a critique.

While the results of the study are not dissimilar from some systematic reviews of technology-based programs (Cheung & Slavin, note small positive improvement due to computer-based reading programs, for instance) (2011), the authors of the study are careful in their analysis to identify their limitations, most notably that the study is meant to be informational. Therefore, they admit that they will not be able to make causal claims about the impact of ClassPace enrollment and use on math skills improvement, that there is potential for confounding variables not included in the dataset or model, and that the study is limited to a single school district (Cook, et al., 2019). Overall, their conclusions stay within these restrictions, without making assertions about the overall impact of the ClassPace program or blended learning. However, it is worth noting just how far from the general population the study sample is: overwhelmingly white (91%), with just 2.9% Latinx, 2.8% multiracial, and 2.8% some other race (ibid). These figures do not match the demographics of the state of Michigan, much less the United States. Fenton, Michigan is a small school district of 3,400 students in just five schools (3 elementary, 1 middle, and 1 high school), which again is not representative of national public schooling. These issues significantly limit the external validity of the study, even if the choice of Fenton, MI was outside of the control of the researchers. Reviewing the publications that Odysseyware puts out, the trend often seems to be that districts and schools opt into the ClassPace program, preventing researchers from randomly selecting a sample or assigning treatment and control groups. It is clear from the report that participation and assignment in the study does not appear to have been random.

There are also some challenges to internal validity, particularly the possibility of contamination. “Two treatment teachers in two different elementary schools were included in the study, along with three comparison teachers from the other elementary school in the district” (Cook, et al., 2019, 2). While contamination is limited because the comparison and treatment teachers are separated from one another, the small size of the district with only three elementary schools suggests that the teachers and students in the district might mix more than they would in a larger district where more schools are spread across a wider geographic area.

 Perhaps the biggest issue in the research stems from the missing data and, from that, challenges in replication of data. As can be seen in the missing values map, there are many missing observations throughout the dataframe, creating Nulls/NAs. According to the researchers, this stems from the merging of two datasets, one containing ClassPace usage information, the other containing student pre- and post-surveys. Most likely, due to the time and financial constraints of the program evaluation, the team was unable to put into place procedures to minimize the missing data from one dataset or the other, or rundown more missing data. Parents had the option of opting their children out of the survey, which contributed to the missing data in the survey dataset and the non-matching student-level data. It is unknown, however, if there was a reason that parents chose to opt their children out that might have created an unseen pattern in that data, influencing the results. As a result of the missing data and mismatched data, it was unclear what conditions were used to create the original sample, and therefore difficult to create an exact match in the sample size of the demographic table, which in the original totaled 123 in the treatment group and 66 in the comparison group, while my version came to 139 and 68, respectively.

Due to differences between the treatment and comparison groups, the researchers were forced to use propensity score matching to align the two groups. Their model for deriving the propensity scores was: *logit(ClassPace enrollment)1 = Intercept + pretest MAP score1 + gender1 + White1 + e1.* This model makes use of what little demographic data is available in the dataframe, as well as the pretest math achievement standardized test score. Based on the qualifications used by the Center for Research and Reform in Education (CRRE) to identify acceptable studies for inclusion in systematic reviews of topics such as elementary reading, math, or social-emotional learning, these are appropriate covariates to include. However, gender and ethnicity are very limited demographic traits. Additional variables, such as socioeconomic status, could have also been useful in determining the differences between groups and creating a more comprehensive model for propensity score matching.

Working with the dataset to recreate certain tables also revealed a few questionable variable and regression model decisions used by the research team, compounded by the absence of certain procedures and models (specifically the student survey subscales regression models focusing on student opinions of math enjoyment, self-efficacy, effort, and answer making sense). This made it difficult to obtain similar results, much less identical ones. One of those procedures that is not given much explanation concerns the study’s use of grand-centered means and regression-adjusted means. While the study’s regression tables included notes about covariates being grand-mean centered, it was unclear which covariates exactly were grand-mean centered. Grand-mean centering is apparently a common practice used by Cook in his studies, but a more appropriate approach may have been to group-center the means at the school level. Similarly, Figure 3 presented the regression-adjusted means for student survey subscales, but as noted above, the models used to calculate the regressions were not provided explicitly. While for some of these tables I was able to recreate the relevant N (all of which were different due to the multiple combined datasets), the figures sometimes varied from what was depicted in the original.

Another issue related to the tables, which simultaneously made it more difficult to replicate and also was a questionable practice, was that Cook and his colleagues used SPSS and Excel to create the tables, hard entering the results of interest. One trend Cook uses in his data is to remove non-significant demographic covariates from the final models for the purpose of parsimony, and to then not display all included covariates and statistics from the tables (such as R2). While this makes the data depicted in the study simpler and easier to understand, it does leave out potentially useful or interesting information that the audience might appreciate or use to better interpret the results. This provides a clear example of how program evaluations are formatted differently from peer-reviewed articles, and the different intended audiences of the two types of papers.

While the authors did a clear job of identifying the limitations of the study that threaten the external validity of the study, the issues described above, including missing data, reliance on grand-centered means, and poor explanation of processes and exclusion of some results are more related to threats to internal validity. The use of grand-centering and the model used for the propensity score, as well as the missing linear model for the association between ClassPace enrollment and student survey subscales raise questions about the quality of the evaluation and its results.

Works Cited

Cheung, A. C. K., & Slavin, R. E. (2011). *The Effectiveness of Education Technology for Enhancing Reading Achievement: A Meta-Analysis*. 48.

Cook, M., Wolf, R., & Eisinger, J. (2019). *Evaluation Study of the ClassPace Program in Fenton, Michigan*. Center for Research and Reform in Education.

LaMalfa, S. (2018). *Elevating the Essentials: Odysseyware Instructional Design Framework for Learning*. Odysseyware.